Image Super-Resolution Using Generative Adversarial Networks with Learned Degradation Operators

Molefe Molefe[1*], *Richard* Klein[2]

[1]School of Computer Science and Applied Mathematics, University of the Witwatersrand, South Africa
[2] School of Computer Science and Applied Mathematics, University of the Witwatersrand, South Africa

Abstract. Image super-resolution is a research endeavour that has gained notoriety in computer vision. The research goal is to increase the spatial dimensions of an image using corresponding low-resolution and high-resolution image pairs to enhance the perceptual quality. The challenge of maintaining such perceptual quality lies in developing appropriate algorithms that learn to reconstruct higher-quality images from their lower-resolution counterparts. Recent methods employ deep learning algorithms to reconstruct textural details prevalent in low-resolution images. Since corresponding image pairs are non-trivial to collect, researchers attempt super-resolution by creating synthetic low-resolution representations of high-resolution images. Unfortunately, such methods employ ineffective downscaling operations to achieve synthetic low-resolution images. These methods fail to generalize well on real-world images that may suffer different degradations. A different angle is offered to solve the task of image super-resolution by investigating the plausibility of learning the degradational operation using generative adversarial networks. A two-stage generative adversarial network along with two architectural variations is proposed to solve the task of real-world super-resolution from low-resolution images with unknown degradations. It is demonstrated that learning to downsample images in a weakly supervised manner is an impactful and viable approach for super-resolution.

1 Introduction

Image super-resolution refers to increasing the resolution of an image while improving the perceptual quality. Completing the task requires corresponding low-resolution and high-resolution image pairs, which are referred to as LR-HR pairs, for supervision. However, a real-world application of super-resolution would not provide corresponding image pairs. Hence, many approaches attempt super-resolution on bicubically downsampled images [5, 11, 23, 26, 32]. Currently, super-resolution techniques fail to attain high-fidelity consistency at a pixel level across the entire generated image. The aim is to compare whether applying state-of-the-art super-resolution techniques on learned degradation performs better than super-resolution on bicubically downsampled images. The investigation primarily focuses on modelling low-resolution images that are later super-resolved. The proposed method generates low-resolution and high-resolution images via a two-stage generative adversarial network. This network learns to degrade high-resolution images to resemble real-

* Molefe Molefe: 1858893@students.wits.ac.za

world degradation and enhance the resultant degradation to improve perceptual quality. Generative adversarial networks learn the distribution of the dataset on which they are trained. As a result, super-resolution generative adversarial networks perform well on test data that have a similar environment to the training data. Hence, it is investigated whether it is plausible to remove the need for pairwise image supervision such that low-resolution counterparts are not required to train super-resolution networks. The first stage of the proposed model is trained on images with unknown degradations using a generative adversarial network to reduce the need for image-pair supervision, effectively reducing the cost of collecting large samples of corresponding image pairs. The second stage of the proposed model is trained on high-resolution images to super-resolve synthetically generated low-resolution images. Thus, the end-to-end pipeline only requires a high-resolution image as input, as opposed to learning to reconstruct relevant perceptual details from low-resolution labels - which are tedious to obtain. This research aims to reduce the cost of requiring large paired data while attaining high perceptual quality from super-resolved images. In doing so, a two-stage GAN that combines a High-To-Low GAN is proposed to learn degradation [3] from low-resolution images with unknown degradations [1] and ESRGAN [30] to super-resolve low-resolution images. Additionally, two variations in the task of image super-resolution are proposed to enhance the perceptual quality of generated images by introducing denoising convolutional neural networks such that one learns to denoise the result of the High-To-Low GAN, and the other learns to denoise the result of the ESRGAN. The results are empirically measured using the similarity between target high-resolution images and super-resolved images using the *Peak-Signal Noise Ratio* and *Structural Similarity* Index. The proposed method performs well perceptually and produces reasonable quantitative results. In summary, the main **contributions** of this paper are:

1. Propose a two-stage generative adversarial network [3], to learn degradation from low-resolution images with unknown degradation.
2. Present an argument for learning to produce better quality degraded images by using denoising model architectures, since generative models for image super-resolution require decent quality low-resolution images.
3. Show that weakly supervised generative models reduce the need for attaining image-pair labels while maintaining image perceptual quality.

2 Related Work

Deep generative models [8] have displayed an impressive ability to generate photorealistic images. Current methods have proposed an unsupervised approach to apply super-resolution on low-resolution images to produce results that downscale correctly [26]. Traditional methods that generate artificial low-resolution images prevent generative models from learning textures present in images and fail to generalise well on real-world examples [3]. Generative models currently present a problem for intelligent image synthesis and image editing. As such, recent papers argue that super-resolution methods that partition the image into overlapped patches and process patches separately ignore the consistency of pixels in overlapped pixels [11].

2.1 Image Upsampling

Image upsampling refers to increasing the spatial resolution of images. Recent methods focus on modelling real-world low-resolution input images. Given the difficulty of data acquisition, researchers aim to find effective ways to improve the reconstruction quality of super-resolution methods. Supervised super-resolution techniques are shown to fail on real-world degradation such as sensor noise and compression artefacts. Zhang et al. [36] investigate such issues and propose frameworks to model real-world degradation. They achieve this by applying various techniques to downgrade the perceptual quality and downsample the spatial resolution. Smaller filter sizes [6] with more mapping layers are argued to increase the speed of super-resolution models that reconstruct the image quality of an image. These methods enable the network to learn upsampling filters directly from the low-resolution representation.

2.1.1 Post-Upsampling

Post-upsampling methods perform convolutional operations and feature extraction on low-resolution representations. They also replace predefined upsampling interpolation-based layers with learnable layers embedded in deep learning models. Feature extraction occurs within the low-resolution representation while upsampling occurs later within the model. Integrating post-upsampling with deep learning improves computational efficiency [7, 28]. However, scale factors are predetermined before training and thus cannot produce multi-scale super-resolution.

2.1.2 Pre-Upsampling

Pre-upsampling methods utilise upsampling algorithms to obtain high-resolution outputs. Deep neural networks refine the resultant feature maps helping models learn an end-to-end mapping from interpolated low-resolution images to high-resolution generated images. These methods introduce artefacts and checkerboard patterns corresponding to image degradation. This occurs because of performing convolutional operations and feature extraction on high-resolution representations [5].

2.2 Image Super-Resolution

Image super-resolution attempts to reconstruct perceptual details from low-resolution images. Deep convolutional networks [5] learn efficient mapping from low-resolution inputs to high-resolution outputs. Dong et al. [5] mention that their proposed SRCNN (Super-Resolution Convolutional Neural Network) can learn super-resolution while coping with three-channel colour images. Deep learning may be necessary for super-resolution at a larger scale [5].

2.2.1 Learning Degradation

Real-world LR images are noisy because of degradations and nuisance factors such as sensor noise, compression artefacts, noise, and blur. These factors are usually unknown and may appear randomly through the pixel space of an image. Ineffective low-resolution models may lead to mediocre performance during test time [3, 37]. Recent techniques [23]

perform bicubic downsampling with anti-aliasing to downsample the spatial resolution of an HR image. Synthetic LR images generated from predefined convolutional operations fail to capture real-world noise. The performance of super-resolution methods degrades because of their inability to effectively model variations of noise in LR images. Hence, generating LR images via downsampling HR images negatively affects the practical applications of super-resolution [4]. Other works combat such a problem by modelling degradation as a combination of several convolutional operations [9]. Gu et al [10] argue that combinatorial degradational models are closer to mimicking real-world degradations.

2.3 Image Denoising

Image denoising describes a set of techniques that remove manifestations of various kinds of noise prevalent in images. Such random noise is caused by signal distortions occurring during image acquisition and transfer over the internet [22]. As a result, images suffer from random degradational factors such as blur, or compression artefacts. The fundamental goal is to reconstruct original images from noisy images. However, researchers recognise the limitations of recent models that perform well numerically but fail to produce visually pleasing results [19]. Literature attempts to improve techniques that remove noise from images to bolster the effectiveness of analysing image observation for different applications, such as medical imaging, photography, remote sensing, and surveillance.

2.4 Convolutional Neural Networks

Convolutional neural networks are neural networks with convolutional operators that have been widely used to solve computer vision problems. They are particularly useful in finding meaningful relationships between neighbouring pixels present in images. Following the success of convolutional neural networks in computer vision tasks [21], state-of-the-art super-resolution techniques leverage CNNs to reconstruct the image quality of a super-resolved image. Consequently, recent deep learning methods leverage the depth of convolutional neural networks by introducing small convolutional filters with a small receptive field to capture orientation [29]. The consensus surrounding deep learning networks is centred around vanishing or saturating nonlinearities. Ioffe and Szegedy [15] discuss a covariate shift which describes the slow training of a deep network due to low learning rates or specific parameter initialisation. Batch normalisation circumvents slow training by normalising layer inputs allowing the models to use higher learning rates for faster convergence. However, accuracy saturation and degradation are consequential with the convergence of a normalised deep network. As a result, He et al. [12] addressed degradation by introducing deep residual frameworks. Fundamentally, the increasing depth increases the number of weights in a network and increases the likelihood of overfitting. Deeply recursive convolutional neural networks address overfitting by repeatedly applying the same convolutional layer. Resultant feature maps from recursive convolutional layers reconstruct pixels in high-resolution images. Skip connections [18] store the input signal during recursions yielding contextual information and allowing the copy of the input signal to be used during target prediction. Generated high-resolution images often lack significant details from their corresponding high-resolution target. Ideally, it would be ideal for super-resolution networks to closely realise the latent details in image quality. Ledig et al. describe SRGAN [23] and note that certain loss functions are responsible for the semantic information within images, such as texture. Furthermore, they believe that MSE (mean squared error) averages the pixel loss between the target and a generated high-resolution resulting in overly smooth image patches.

2.5 Generative Adversarial Networks

The zero-sum game is a game in which one player's gain is another's loss. Games, like chess, characterise such a situation where the net benefit for individual players is eventually zero. A generative adversarial network is an architecture [8] that suggests two models which compete against each other. The generative model, G, is trained to generate the data distribution, and the discriminator D must determine if a sample belongs to the data distribution generated by G or training data. In other words, the generator attempts to fool the discriminator that the sample belongs to in the training data. The adversarial loss function for training generative models is formulated as

$$\min_G \max_D \mathbb{E}_{x \in X}[\log \mathbf{D(x)}] + \mathbb{E}_{z \in Z}[log(1 - D(G(z)))] \quad (1)$$

where *x* is sampled from the original data and *z* is a sampled noise vector.
The generator learns a generator distribution by building a mapping function from a prior noise distribution to data space while the discriminator returns the probability that the input belonged to the training data.

2.5.1 Super-Resolution GAN

Ledig et al. [23] proposed a *super-resolution GAN* that increases the spatial resolution of an image while improving the perceptual quality. They argued that previous loss functions fell short of recognizing subjective qualities in pixels. Hence, the paper introduces a perceptual loss function which consists of *adversarial loss* and *content loss*. Authors borrowed the notion of residual networks with skip connections [13] to address vanishing gradients in their model.

2.6 Loss Functions

2.6.1 Pixel-Wise Loss

Per-pixel loss measures the pixel-wise difference between output and target images. Additionally, minimising the mean squared error [25] finds the pixel-wise averages of plausible solutions resulting in overly smooth details in pixels with textual context.

2.6.2 Perceptual Loss

Johnson et al. [16] argue that per-pixel loss does not capture perceptual differences between output and target images. Optimising perceptual loss, based on high-level features extracted from pre-trained networks, generates high-quality images. Specifically, papers define perceptual losses to measure high-level perceptual and semantic differences between output and target images. Style transfer methods used style reconstruction loss to penalise differences in style while using feature reconstruction loss to penalise the output when it deviates from the content in the original target.

2.7 Image Quality Assessment

Image quality assessment refers to the qualitative and quantitative measurements of super-resolution methods. The model considers the objective methods to assess the quality of the results.

2.7.1 Structural Similarity

The structural similarity index measures (SSIM) the similarity between two images. It is useful when comparing the ground truth image and image generated by super-resolution algorithms. The following equations describe the SSIM as [31]:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \qquad (2)$$

Where *x, y* represent the images that are being compared respectively, and c represent the variables that stabilize the division with a weak denominator.

2.7.2 Peak Signal-To-Noise Ratio

A well-used metric used for image quality is the peak signal-to-noise ratio. It is defined as the ratio between the maximum value of a signal and the power of distorting noise that affects the quality of its representation. The following describes its mathematical formulation [31]:

$$PSNR = 10.\log_{10}\cdot\frac{L^2}{\frac{1}{N}\sum_{i=1}^{N}(I(i) - \hat{I}(i))^2} \qquad (3)$$

where L equals 255 in general cases using 8-bit representations, N represents the number of pixels, both the I's represent the input image and generated image respectively. Ledig et al. [23] set the SRGAN with high upscaling factors, up to four measured by the PSNR. The PSNR, measured in decibels, struggles to assess the quality of the image and does not correlate well with human judgement. There is evidence of super-resolution methods that record high PSNR values corresponding with images that result in overly smooth patches [23].

3 Method

In this section, the methodology that implements the proposed model is discussed.

3.1 Overview

The goal of learned degradation is to investigate its effectiveness in the practical application of super-resolution. The approach uses the High-To-Low GAN [3] to learn degradation from images with unknown degradation and super-resolve the result using the ESRGAN architecture [30]. The generative model that accomplishes the task of image super-resolution learns to generate SR images from synthetic low-resolution images. The following equations regarding downsampling and upsampling generators are defined as:

$$\hat{I}_{LR} = \mathbb{G}_{\downarrow}(I_{HR}; \theta_{LR}) \tag{4}$$

$$\hat{I}_{SR} = \mathbb{G}_{\uparrow}(\hat{I}_{LR}; \theta_{HR}) \tag{5}$$

where *LR* and *SR* represent generated low-resolution and super-resolved images respectively, and G refers to a generator with parameters θ.
The motive is to investigate the effectiveness of removing noise from learned low-resolution images by moderating the relevance of learned degradation. The investigation is approached by proposing a denoising architecture that learns to remove noise from degraded images. The denoiser architecture is either applied to the result of a High-To-Low GAN to denoise the synthetic low-resolution image or applied to the result of the ESRGAN to remove image artefacts from high-resolution pixels. The following equations describe how to denoise low-resolution and high-resolution images respectively:

$$\hat{N}_{LR} = \mathbb{N}(\hat{I}_{LR}; \theta_{NR}) \tag{6}$$

$$\hat{N}_{SR} = \mathbb{N}(\hat{I}_{SR}; \theta_{NR}) \tag{7}$$

where N refers to a model that denoises the input images with parameters θ.

3.2 Image Super-Resolution

Super-Resolution describes an image upsampling technique that increases the perceptual quality of a low-resolution image. The trained models perform image-to-image translation by mapping low-resolution images to high-resolution images. Supervised super-resolution depends on the acquisition of LR-HR pairs. Effective techniques are investigated to explore various methods of modelling real-world low- resolution images to avoid taking photographs of LR-HR image pairs. Common interpolation-based downsampling techniques, such as bicubic downsampling, smoothen the texture details in high-resolution images, disallowing super-resolution models from representing original texture details in real-world low-resolution images. The goal is to model a real-world low-resolution input to improve the generalisability of a super-resolution model. Super-resolving bicubically downsampled images present artefacts and undesirable details in generated HR im- ages. Consequently, the focus is specifically on improving the perceptual quality rather than simply increasing the spatial resolution of a low-resolution image. Hence, a common technique is to downsample the spatial resolution followed by quality degradation to generate corresponding LR images from HR images [31].

$$\hat{I}_{LR} = \mathbb{D}(I_{HR}, \theta_{LR}) = (I_{HR} * k) + \eta \tag{8}$$

where D applies a downsampling technique by applying convolution k on an image. η represents additive noise that may be used to incorporate additional information on low-resolution pixels. Previous methods that approach super-resolution using LR-HR pairs heavily depend on supervision. Obtaining LR-HR pairs is an expensive task. Data collection requires photographing both image pairs under exact circumstances. Ideally, it would be preferable for general models to upsample images without corresponding low-resolution representation.

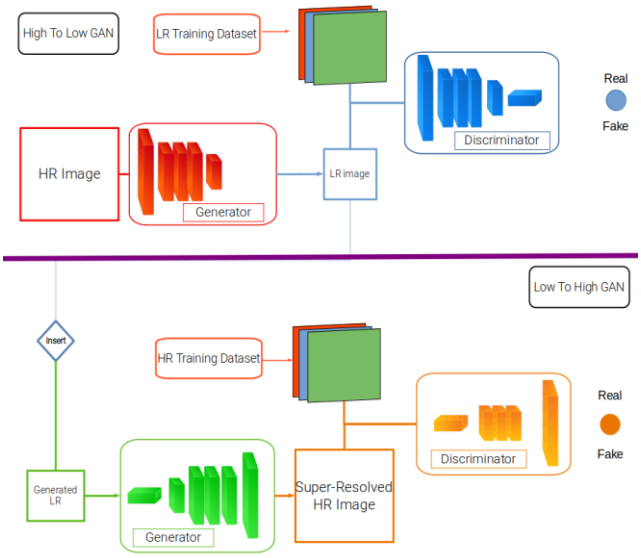

Fig. 1. Proposed Architecture

3.3 Two-Stage Adversarial Architecture

Fig. 1 demonstrates a two-stage GAN [3] to allow super-resolution methods to learn nuisance factors from real-world degradations. This is achieved by using a generative model to learn degradational operators from low-resolution images with unknown degradation. The **DIV2K** dataset [1] collected high-resolution images used for the CVPR 2017 challenge and provided corresponding low-resolution images that were synthetically generated with degradation operators unknown to the participants. The goal of the proposed architecture is to reduce the cost of collecting image pairs required for training super-resolution images. Hence, in Fig. 1, a weakly supervised degradation model is proposed to learn image downsampling to generate a corresponding paired dataset for the super-resolution task. The resultant low-resolution images are directly trained to produce corresponding high-resolution images using generative adversarial networks. The *High-To-Low* generative model aims to capture the distribution of different LR representations while the *Low-To-High* generative model learns to generate pixels with pleasing visual details. This hypothesis is that sampling from an LR generative model will be closer to real-world noise than bicubically downsampled images with anti-

aliasing. Additionally, a denoising CNN model is added [35] to learn how to denoise either generated low-resolution images or high-resolution images.

3.3.1 High-To-Low GAN

A well-used metric used The *High-To-Low GAN* is tasked to learn how to generate LR images. The generated LR images will serve as inputs to the *Low-To-High* network to train a super-resolution model. Given that LR images are subject to unknown degradations, the proposed model borrows the learning degradation architecture that learns to increase the quality of real-world images corrupted by noise and artefacts. It models a variety of image degradational noise by learning a probability distribution that samples LR images.

3.3.2 Low-To-High GAN

The low-to-high GAN is responsible for generating HR images with realistic textures. In particular, the model follows an ESRGAN architecture [30] that improves the perceptual loss by using pre-trained features before activation. Researchers [30] argue that their model achieves a more pleasing quality with realistic textures in comparison to the SRGAN [23] paper.

3.4 End-To-End Two-Stage Denoising Architecture

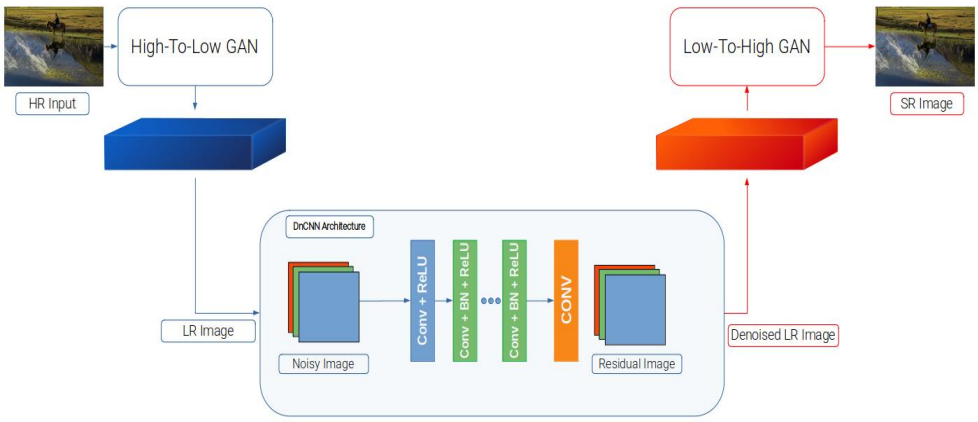

Fig. 2. End-to-end network architecture

In the context of image super-resolution, the goal remains to enhance the spatial resolution of a low-resolution image whilst suppressing the noise without losing relevant perceptual details. The architecture given in Fig. 2 aims to resolve residual artefacts created by generative networks and decrease noise prevalent in synthetic images. The pipeline is trained in an end-to-end manner where the input of the network is a high-resolution image that is downsampled, denoised and further upsampled for the task of super-resolution. The architecture mimics a denoising autoencoder that learns to create an identity mapping.

However, the fundamental purpose of the degradational generative model is to push the super-resolution generator to learn to upsample images while suppressing the noise present in the resultant super-resolved images.

3.5 Loss Functions

The aim is to effectively super-resolve LR images by using loss functions measuring reconstruction error and guiding the super-resolution models' optimization. SRGAN authors [23] proposed a loss function that takes advantage of pre-trained models to infer texture in image details. The ESRGAN further extends the adversarial loss to incorporate texture information in the adversarial training of the generator. Through adequate training, the generator will produce outputs sampled from a distribution learned from adversarial training

3.5.1 High-To-Low GAN

The *High-To-Low* GAN is comprised of a suite of loss functions that work together to form a cohesive learning mechanism.

Pixel Loss. Pixel loss measures the pixel-wise distance between two images

$$l_{pixel} = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H}(F(I^{HR})_{i,j} - G_{\theta_G}(I^{d})_{i,j})^2 \qquad (9)$$

Hinge Loss. Generative adversarial networks are notorious for training instability, often caused by random noise generation, mode collapse, and vanishing gradient. Recent methods [27] use spectral normalisation to tackle such problems presented by adversarial training using GANs.

$$l_{GAN} = \mathbb{E}_{x\sim\mathbb{P}_r}[\min(0,-1+D(x))] + \mathbb{E}_{\hat{x}\sim\mathbb{P}_g}[\min(0,-1-D(\hat{x}))] \qquad (10)$$

where P_r is the data distribution and P_g is the generator G distribution defined by $x\hat{} = G(x)$.

High-To-Low Loss. Authors [3] argue that a GAN-centred approach motivates the GAN to drive the image generation process. As a result, L_2 pixel loss is responsible for refining LR images to make them sharper for super-resolution

$$l = \alpha l_{pixel} + \beta l_{GAN} \qquad (11)$$

3.5.2 Low-To-High GAN

The *Low-To-High* GAN is comprised of loss functions that learn to reconstruct HR pixels and minimize the effect of blurry artefacts.

Perceptual Loss. The perceptual loss constrains the pre-trained model features before activation to model perceptual similarity between real HR images and super-resolved images.

$$L_G = L_{percep} + \alpha L_G^{Ra} + \eta L_1 \qquad (12)$$

where $L_1 = E_{xi} ||G(x_i)-y||_1$ is the content loss that evaluates distance between generated super-resolved image and real HR image. Researchers [30] argue that perceptual loss will pre- vent the drawbacks presented in the [23] SRGAN paper

Adversarial Loss. The ESRGAN employs a relativistic GAN [17] to predict the probability that real images are relatively more realistic than images generated from the fake dataset. Researchers argue relativistic GANs from a super-resolution perspective recover more detailed textures. The generator adversarial loss is defined as:

$$L_G^{R_a} = -\mathbb{E}_{x_r}[log(1 - D_{R_a}(x_r, x_f))] - \mathbb{E}_{x_f}[log(D_{R_a}(x_f, x_r))] \qquad (13)$$

While the discriminator loss is defined as:

$$L_D^{R_a} = -\mathbb{E}_{x_r}[log(D_{R_a}(x_r, x_f))] - \mathbb{E}_{x_f}[log(1 - D_{R_a}(x_f, x_r))] \qquad (14)$$

3.5.3 DNCNN Denoising Architecture

Average Mean Squared Error. The error is measured between the desired residual images and estimated outputs from the noisy input

$$l(\theta) = \frac{1}{2N}\sum_{i=1}^{N} ||\mathcal{R}(y_i;\theta) - (y_i - x_i)||_F^2 \qquad (15)$$

4 Experiments

In this section, there is a discussion on how the dataset is obtained and split to allow effective training.

4.1 Dataset

The two-stage GAN is trained on the DIVerse 2K resolution image dataset taken from the NTIRE 2017 challenge on *single image super-resolution* [1]. The data was collected from the internet and contains 1000 images with very high resolution. Images are degraded via two methods.

- **Standard:** Bicubic downsampling with anti-aliasing
- **Unknown Downscaling Operations:** Diverse set of downsampling techniques with additive noise to model real-world degradations

4.2 Train-Validation-Test Split

- **Training Data:** 800 high-resolution images
- **Validation Data:** 100 high-resolution images
- **Testing Data:** 100 high-resolution images

The process required 800 training images from the DIVerse 2K dataset to train the two-stage GAN to super-resolve images. The validation dataset aided us in fine-tuning the model by selecting the best hyperparameters. The testing data is used to evaluate and report the quality of the model using image quality assessment techniques

4.3 Data Preparation

Image Size. High-Resolution images have high perceptual detail and size. Such images present a problem for training expansive models. Hence, only random crops of size 256 × 256 for each LR-HR pair from the dataset are considered to alleviate pressure on the GPU.

4.4 Training

In the following section, there is a discussion on what is required for the training process to perform weakly supervised super-resolution.

4.4.1 Overview of the Training Process

The *High-To-Low GAN* is responsible for generating synthetic low-resolution images. A dataset of high-resolution images that have corresponding low-resolution counterparts with unknown degradations is collected to perform the experiments. While doing so, the experiment required collectively selecting random crops of corresponding LR-HR image pairs to train the large networks. The synthetic LR images effectively replace the original supervised corresponding LR images in the process of learning how to upsample images. Thus, this model utilizes weak supervision to train the super-resolution model from artificial low-resolution images. The result is directly fed into the ESRGAN architecture to generate corresponding high-resolution images. The models are both trained with different hyperparameters to achieve satisfactory results. The weakly-supervised training allows our method to perform well independent of original supervised image pairs.

4.4.2 Hyperparameters

The hyperparameters chosen for the *High-To-Low GAN* are listed as follows [3]:

- **Learning Rate:** $1e^{-4}$
- **Epochs:** 1409
- **High-To-Low Loss:** $\alpha = 1$, $\beta = 0.05$, (12)
- **Adam Optimization:** $\beta_1 = 0$, $\beta_2 = 0.9$

The generative model (ResNet) is trained before training the GAN. The resultant ResNet model is further fine-tuned using an adversarial approach to learn texture. The hyperparameters chosen for the ESRGAN architecture are listed as follows [30]:

- **Learning Rate**: $1e^{-4}$
- **Epochs:** 1409
- **Perceptual Loss:** $\lambda = 5 \times 10^{-3}$, $\eta = 1 \times 10^{-2}$ (13)
- **ADAM Optimization:** $\beta_1 = 0.9$, $\beta_2 = 0.999$

The DNCNN architecture used to denoise the resultant low-resolution image or high-resolution image has the following hyperparameters:

- **Learning Rate:** $1e^{-4}$
- **Epochs:** 1409
- **Adam Optimization:** $\beta_1 = 0.9$, $\beta_2 = 0.999$

4.5 Results

The proposed two-stage models that super-resolve low-resolution images, given in Fig 2, during evaluation accurately reconstruct the ground-truth high-resolution image. However, the vanilla two-stage model without the denoising architecture outperforms the end-to-end pipeline methods that include denoising to reduce noise on either side. The proposed two-staged model, taken from Fig. 2, trains a model that learns a low-resolution representation that is super-resolved later in the ESRGAN generator.

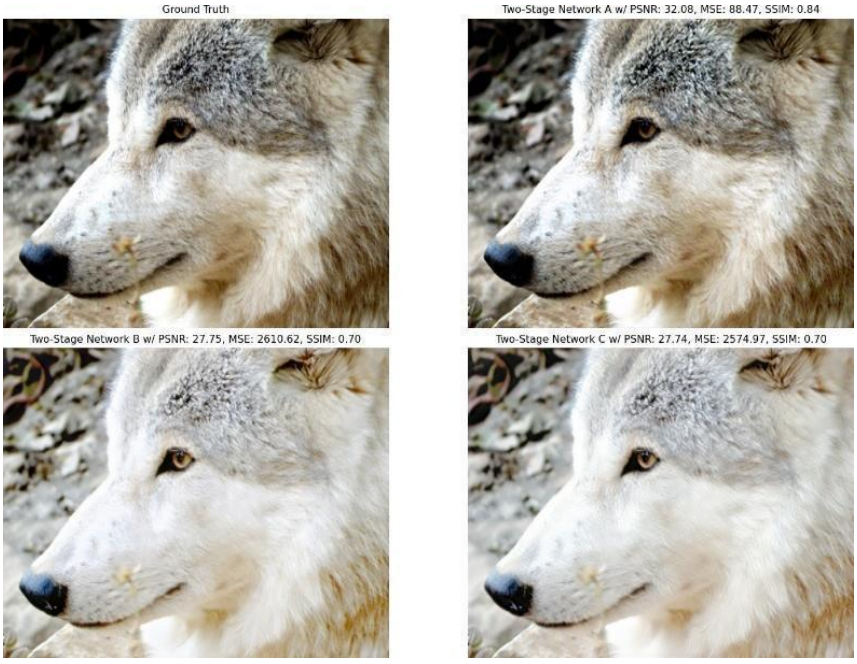

Fig. 3. Top-left (Ground Truth), Top-right (Two-Stage Network A) - without denoising architecture, Bottom-left (Two-Stage Network B) -denoise super-resolved image, Bottom-right (Two-Stage Network C) - denoise synthetic LR image

Dataset	SRGAN	ESRRestNet	ESRGAN	ESRGAN w/o Gradient Penalty	Two-Stage Network A (Ours)	Two-Stage Network B (Ours)	Two-Stage Network C (Ours)
Set5 [2]							
PSNR (dB)	19.86 (1.28)	25.15 (2.67)	24.40 (1.40)	**27.83** (2.51)	19.54 (2.78)	15.05 (1.31)	13.91 (0.84)
SSIM	0.59 (0.10)	0.70 (0.15)	0.75 (0.11)	**0.80** (0.10)	0.56 (0.13)	0.57 (0.12)	0.41 (0.10)
Set14 [34]							
PSNR (dB)	20.11 (1.91)	**27.61** (2.49)	22.89 (2.48)	25.72 (2.88)	19.20 (1.83)	15.77 (2.64)	14.30 (2.07)
SSIM	0.55 (0.13)	**0.80** (0.09)	0.66 (0.16)	0.70 (0.15)	0.59 (0.16)	0.53 (0.17)	0.42 (0.16)
BSD100 [24]							
PSNR (dB)	19.84 (2.14)	23.06 (2.96)	22.14 (2.91)	**23.24** (2.96)	17.93 (2.13)	16.32 (2.74)	14.91 (2.07)
SSIM	0.53 (0.10)	0.66 (0.13)	0.64 (0.13)	**0.66** (0.13)	0.63 (0.11)	0.51 (0.12)	0.40 (0.13)

Table 1: Comparison of SRGAN [23], ESRGAN [30], and the proposed two-stage GAN architectures. The table stores the PNSR (with standard deviation) values measured in decibels and the SSIM (with standard deviation) that measures the similarity between the produced output and target image

Quantitatively, the ESRGAN trained without a gradient penalty outperforms the rest of the methods in terms of PSNR and SSIM (refer to Table 1). The results were generated by running each of the models on individual test datasets [2, 24, 34] and averaged the image quality assessment recordings.

4.5.1 High-To-Low GAN

In theory, generating LR images that capture variations resembling real-world noise is plausible. However, generated HR images produce noise in the dark regions of an image. The image dynamic range is stretched further resulting in overwhelming dark patches in regions where the original image was dark. After investigating the result of the *High-To-Low GAN*, the hypothesis is that since LR images have smaller spatial dimensions, generated LR images are prone to artefacts and pixels that erroneously summarise contextual information. Implying that noise present in LR images will also be super-resolved thus magnifying the presence of noise in generated HR images

4.5.2 Misrepresenting Colour

This method emphasises specific colours on patches that have densely packed similar pixels. It shows that the model clumsily super-resolves low-resolution patches that have an affinity for dark neighbourhoods. Fig. 4 demonstrates that the model assumed that all patches that had low lighting had a purple colour. This is due to the *High-To-Low* GAN learning from lacking information. As a result, *Low-To-High* GAN learns a noisy distribution and super-resolves patches that are surrounded by significantly dark pixels

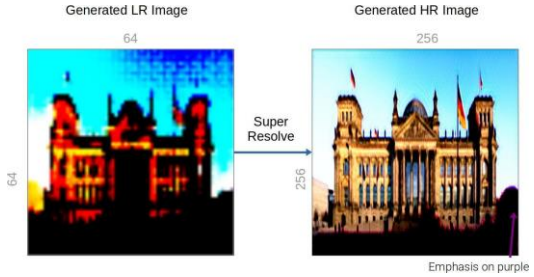

Fig. 4. Demonstrating the impact of generative adversarial networks on low-spatial information

4.5.3 Poor Denoiser

A denoising architecture is initiated to alleviate the drawbacks created by the *High-To-Low* GAN. The lighter regions of the image, given in Fig. 4, generated by the end-to-end denoising architecture, given in Fig. 2, appear to be emphasised. Resultant images demonstrate a colour space that is more luminant and bright in comparison to the other methods. Although their structural similarity seems low, these results may be desirable in a real-world context. Generally, sought models are those that generalise well on unseen data. Consider LR datasets without corresponding HR counterparts. Images with a broader colour space may be more desirable in terms of perceptual quality. Denoising architectures add extra complexity such that it becomes harder to fine-tune hyperparameters for models to converge. As a result, finding conclusive evidence of whether the model performs well becomes an open-ended problem. Fig. 5 attempts to visualise the effect of denoising the result of synthetic low-resolution reconstructed by the *High-to-Low* GAN. The figure also demonstrates the lack of representative pixels created by the generative model that downsamples high-resolution images. This explains the light textures, shown in Fig. 5, presented by the end-to-end pipeline that learns to super-resolve from denoised low-resolution images. Fig. 4

demonstrates that denoising architectures convincingly learn how to illuminate pixels that were inherently dark in the training set. However, the quantitative measurements, presented in Table 1 as *Two-Stage Network C*, show that poor denoising architectures result in low PSNR values that measure the closeness between ground truth and prediction. There is more work to be done to learn how to appropriately tweak denoising architectures to learn how to remove arbitrary noise and artefacts prevalent in real-world images. There is little compelling evidence that there is any difference between learning to denoise after generating LR images or after HR images.

F**ig. 5.** Demonstrating the effect of a poor denoiser that disallows the super-resolution model to upsample images accurately

5 Future Work and Discussion

In this section, there is a discussion on the impact of super-resolution on the real-world and how researchers may tackle the problem moving forward.

5.1 Super-resolution as an ill-posed problem

Menon et al [26] showed that low-resolution images correspond to multiple high-resolution images. Super-resolution is fundamentally ill-posed as a mapping from a low-resolution input results in multiple outputs. Each high-resolution output could correctly represent the reconstruction of a low-resolution image. Hence, subjectivity thus becomes a determining factor for perceptual quality. Such idiosyncrasies allow for vast interpretations that are contextual for each scenario.

5.2 Super-resolution as an ill-posed problem

Reconstructing texture and high-quality details in images have gained notoriety in computer vision research. Recent methods have proposed deep and dense networks to understand the relationships between pixels in images. They have also considered optimization techniques to help deep networks converge faster. As a result, learning feature extraction from low-resolution representations has been neglected. The hypothesis is that learning how to reduce noise in low- resolution inputs should teach models to achieve pleasing results for super-resolution using learned degradation. Authors [14] proposed a variational autoencoder with a training criterion that learns flexible posterior distributions. They observed that their method can help improve the performance of variational autoencoders. This method can be applied to denoise the degradation in low-resolution images. Thus, allowing super-resolution models to produce results with pleasing textures

5.3 Practical Applications of Super-Resolution

5.3.1 Surveillance

Surveillance technology aims to monitor and observe continuous and ongoing activities. Practical applications of surveillance, such as security, require acquiring high-quality photographic information to make meaningful decisions. These methods often require large and expensive equipment to photograph incidents. Image super-resolution can be used to magnify the noisy details in recorded footage [33].

5.3.2 Medical Imaging

Medical institutions capture medical images using a medium that engages an elementary particle (or wave) with the tissue. Technologies that capture the medical photograph suffer from signal limitations such as attenuation and scattering. Spontaneous tissue rhythms, such as a heartbeat or inhalation, may misrepresent the resolution of a medical image. Hence, super-resolution methods may enhance the perceivability of medical images such as ultrasounds [20].

6 Conclusion

Super-resolution is an emerging field in computer vision that solves many practical problems in enhancing the spatial dimensions of an image. The progression of technology has witnessed the storage of different mediums since its inception. However, the increased computation implies that society will struggle to interpret information from the past. Super-resolution allows the preservation of historic information using deep learning. Furthermore, it serves as a cheaper alternative to produce high-fidelity content without expensive equipment. Current techniques fail to generalise on unseen data such that super-resolution produces blurry artefacts. Hence, researchers have deployed different techniques to learn how to enhance images resembling real-world degradation. The proposed model has demonstrated the extent of applying learned degradation as a method to learn super-resolution. Although ESRGAN quantitatively performs better while using corresponding image pairs, the proposed approach achieves reasonable metric scores without original supervised image pairs demonstrating the potential of investigating weakly supervised super-resolution techniques. Furthermore, the experiments have demonstrated the potential of using weakly supervised techniques to train super-resolution models. The proposed method greatly reduces the cost of acquiring supervised image pairs. Additionally, the results have highlighted that super-resolved images from learned degradation suffer from low-resolution noise. A promising endeavour for approaching super-resolution using learned degradation should consider employing dense networks to extract features from low-resolution images more effectively

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super- resolution: Dataset and study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1122–1131, 2017.

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low- complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.

[3] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a GAN to learn how to do image degradation first. *CoRR*, abs/1807.11458, 2018.

[4] Honggang Chen, Xiaohai He, Linbo Qing, Yuanyuan Wu, Chao Ren, Ray E. Sheriff, and Ce Zhu. Real-world single image super-resolution: A brief review. *Information Fusion*, 79:124–145, 2022.

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks, 2015.

[6] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution con- volutional neural network, 2016.

[7] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution con- volutional neural network. *CoRR*, abs/1608.00367, 2016.

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sher- jil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[9] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. *CoRR*, abs/1904.03377, 2019.

[10] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction, 2019.

[11] Shuhang Gu, Wangmeng Zuo, Qi Xie, Deyu Meng, Xiangchu Feng, and Lei Zhang. Convo- lutional sparse coding for image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[14] Daniel Jiwoong Im, Sungjin Ahn, Roland Memisevic, and Yoshua Bengio. Denoising cri- terion for variational auto-encoding framework, 2016.

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network train- ing by reducing internal covariate shift, 2015.

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016.

[17] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan, 2018.

[18] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution, 2016.

[19] Claude Knaus and Matthias Zwicker. Progressive image denoising. *Image Processing, IEEE Transactions on*, 23:3114–3125, 07 2014.

[20] D. Kouame and M. Ploquin. Super-resolution in medical imaging: An illustrative approach through ultrasound. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 249–252, 2009.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher

J. C. Burges, L´eon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing*
Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States, pages 1106–1114, 2012.

M. Lebrun, M. Colom, A. Buades, and J. M. Morel. Secrets of image denoising cuisine.
Acta Numerica, 21:475–576, 2012.

[23] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Ale- jandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. 2017.

[24] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001.

[25] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error, 2016.

[26] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models, 2020.

[27] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral nor- malization for generative adversarial networks. *CoRR*, abs/1802.05957, 2018.

[28] Wenzhe Shi, Jose Caballero, Ferenc Husz´ar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *CoRR*, abs/1609.05158, 2016.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[30] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. ESRGAN: enhanced super-resolution generative adversarial networks. *CoRR*, abs/1809.00219, 2018.

[31] Zhihao Wang, Jian Chen, and Steven C. H. Hoi. Deep learning for image super-resolution: A survey. 2020.

[32] Jiahui Yu, Yuchen Fan, Jianchao Yang, Ning Xu, Zhaowen Wang, Xinchao Wang, and Thomas S. Huang. Wide activation for efficient and accurate image super-resolution. *CoRR*, abs/1808.08718, 2018.

[33] Linwei Yue, Huanfeng Shen, Jie Li, Qiangqiang Yuan, Hongyan Zhang, and Liangpei Zhang. Image super-resolution: The techniques, applications, and future. *Signal Processing*, 128:389–408, 2016.

[34] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse- representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010.

[35] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *CoRR*, abs/1608.03981, 2016.

[36] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations, 2018

[37] Shaolei Zhang, Guangyuan Fu, Hongqiao Wang, and Yuqing Zhao. Degradation learn- ing for unsupervised hyperspectral image super-resolution based on generative adversarial network. *Signal, Image and Video Processing*, 15:1–9, 11 2021.